PHIL 7001: Midterm Exam

Professor Boris Babic

Instructions:

This is the midterm examination for PHIL 7001. The assignment will be graded out of a total of 100 points. It is closed book, but you can have one double spaced cheat sheet, on which you can write anything you like.

You have 180 minutes to complete this test. If you finish early, you may submit your answers to the test invigilators and leave early. However, in order to avoid too much commotion, during the last 30 minutes please do not leave. That is, if you finish with less than 30 minutes remaining, please remain at your desk quietly until the exam is over.

Please make sure to write as clearly and legibly as possible. This will make it much easier for grading purposes.

You can write your answer in the exam booklet itself. Directly below each question. If you need more space, you can use the back side of each paper of the exam booklet.

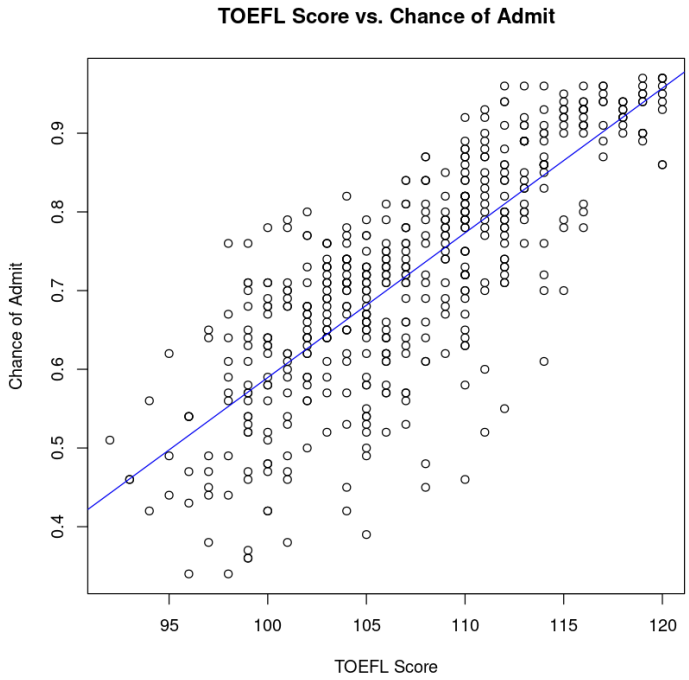
Calculators are allowed, and you can use any calculator you wish. Smart devices of any kind are not allowed.

Please write your name and student number clearly on the first page of your submission.

There should be no trick questions on the exam. If anything seems unclear or ambiguous, then you can use your judgment to resolve the ambiguity, and note that you have done so in your answer. It is always best to state your assumptions as clearly as possible.

Problem 1: Fundamentals of R (20 points)

1A (5 Points). Below is a plot generated from a dataset we saw in our class exercise (the TOEFL data).

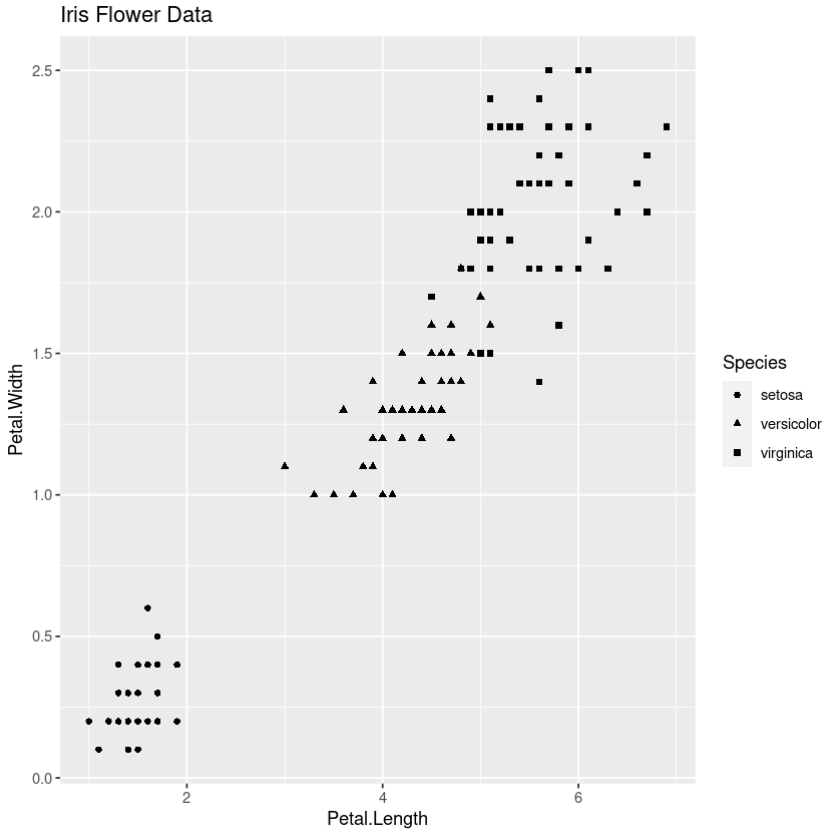


Please complete the blanks in the following code, which was used to generate the above plot.

```
plot(admissions_data$`TOEFL Score`, admissions_data$`Chance of Admit`,
     xlab = "TOEFL Score",
     ylab = "Chance of Admit",
     main = "TOEFL Score vs. Chance of Admit"
    )

abline(model, col = "blue")
```

1B (5 Points). Below is a plot generated from a well-known pedagogical dataset called the Iris flower data.



Please complete the blanks in the following code, which was used to generate the above plot.

```
ggplot(iris, aes(Petal.Length, Petal.Width)) +
    geom_point(aes(shape = Species)) +
    ggtitle("Iris Flower Data")
```

1C (5 Points). Which species has the shortest petals, in terms of both length and width?

setosa

1D (5 Points). Which species has the longest petals, in terms of both length and width?

virginica

Problem 2: Probability with R (20 Points)

2A (5 Points). Suppose that we have a financial portfolio made up of 100 separate equity (stock) investments. For each stock, we have measured its performance increase over the past year. We find that the average increase is 12 percent, with a standard deviation of 4 percent. If you assume that the stock performances are normally distributed, you can calculate the probability that a randomly selected stock increased by at least 20 percent. To do this, please complete the R code below.

```
pnorm(0.2, mean = 0.12, sd = 0.04, lower.tail=FALSE)
```

or

```
pnorm(20, mean = 12, sd = 4, lower.tail=FALSE)
```

Note: pnorm(20%, mean = 12%, sd = 4%, lower.tail=FALSE) is incorrect and will result in syntax error when executing in R.

2B (5 Points).

Suppose you wish to find the probability that a randomly selected stock increased between 10 percent and 20 percent. Please write below the R code that you would use to compute this quantity. Hint: please be careful with your choice for lower.tail.

```
pnorm(0.2, mean = 0.12, sd = 0.04) -
pnorm(0.1, mean = 0.12, sd = 0.04)
```

or

```
pnorm(0.1, mean = 0.12, sd = 0.04, lower.tail=FALSE) -
pnorm(0.2, mean = 0.12, sd = 0.04, lower.tail=FALSE)
```

2C (5 Points).

In a certain population, 20 percent of people have a particular genetic trait. If you randomly sample 200 people, what is the probability that fewer than 30 of them have that trait? Please write below the R code that you would use to compute this quantity

```
pbinom(29, size = 200, prob = 0.20)
```

2D (5 Points).

Consider the following function.

```
sum(dbinom(0 : 3, size = 25, prob = 0.55))
```

Please describe in one to three sentences what kind of probability this function will produce. Hint, for example, your answer may say: this function is generating the probability of observing BLANK from a BLANK distribution with BLANK parameters.

This function is generating the probability of observing **at most 3** successes out of **25 trials** from a **Binomial** distribution with a **success probability of 0.55**.

Problem 3: Bayes' Rule (30 Points)

Suppose that 2% of people in the United States have Cystic Fibrosis. Suppose that a test has been developed to check whether or not someone has this disease.

The test has a sensitivity rate of 0.92. Hint: you should be able to express this in terms of the appropriate probability.

The test has a specificity rate of 0.98. Hint: you should be able to express this in terms of the appropriate probability as well.

3A (5 Points).

The false positive rate can be given as 1 - specificity(0.98) = 0.02.

3B (5 Points).

The false negative rate can be given as 1 - sensitivity(0.92) = 0.08.

3C (15 Points).

Find the probability that a patient has Cystic Fibrosis, given that they received a positive test. That is, find:
Pr(Patient has Cystic Fibrosis | Positive Test).

$$\begin{aligned} Pr(\text{Positive Test}) &= (\text{Sensitivity} \cdot \text{CF}) + (\text{False Positive Rate} \cdot \text{Non-CF}) \\ &= (0.92 \cdot 0.02) + (0.02 \cdot 0.98) = 0.0184 + 0.0196 = 0.038 \end{aligned}$$

$$\begin{aligned} Pr(\text{CF} \mid \text{Positive Test}) &= \frac{Pr(\text{Positive Test} \mid \text{CF}) \cdot Pr(\text{CF})}{Pr(\text{Positive Test})} \\ &= \frac{0.92 \cdot 0.02}{0.038} \\ &\approx 0.4842 \text{ (or } 48.42\%) \end{aligned}$$

3D (5 Points). Suppose that you now wanted to compute the probability that a patient has Cystic Fibrosis, given that they tested positive for Cystic Fibrosis not just once, but two times in a row. You can assume the tests are independent of each other. You do not have to actually calculate this, rather, I want you to answer the following: Which quantity would you now use as your base rate? Hint: in part C, you used 2% as your base rate. Now, the base rate should change. What should it be?

The updated base rate after getting one positive test result would be 48.42%.

Problem 4: Linear Regression (30 Points)

Consider the following well-known pedagogical dataset in R, called Cars.

```
head(cars, n=10)
```

```
##    speed dist
## 1      4    2
## 2      4   10
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
## 8     10   26
## 9     10   34
## 10    11   17
```

The cars dataset gives Speed and Stopping Distances of Cars. This dataset is a data frame with 50 rows and 2 variables. The rows refer to cars and the variables refer to speed (the numeric Speed in mph) and dist (the numeric stopping distance in ft.)

4A (5 Points).

You want to build a linear (regression) model, whereby you model a car's stopping distance as a function of its speed. That is,

$$\text{Stopping Distance} = \hat{\beta}_0 + \hat{\beta}_1 \text{Speed}.$$

To do this, please complete the blanks below.

```
lm(dist ~ speed, data = cars)
```

4B (5 Points).

If you were to run the model in 4A, above, you would obtain the following estimates.

$$\hat{\beta}_0 = -17.579,\ \hat{\beta}_1 = 3.932.$$

Please explain in 2-3 sentences the meaning of $\hat{\beta}_0$ and of $\hat{\beta}_1$.

In this context, $\hat{\beta}_0 = -17.579$ is the **estimated** intercept, which is the predicted stopping distance when the speed of the car is zero. $\hat{\beta}_1$ represents the **estimated** coefficient for the predictor "speed". In this case, $\hat{\beta}_1 = 3.932$ indicates that for each unit increase in speed, the predicted stopping distance is expected to increase by approximately 3.932 units.

Note: It's important to understand that $\hat{\beta}_i$ is different from β_i, where β_i is the **parameter** and $\hat{\beta}_i$ is the **estimated** statistic of β_i

4C (5 Points).

The following is the fully summary of the regression model, above.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

First, please explain in 2-3 sentences the general meaning of p-value and how it relates to hypothesis testing.

A p-value measures the probability of observing a result as extreme as, or more extreme than the observed results, assuming that the null hypothesis is true. In hypothesis testing, the p-value serves as a statistical measure that quantifies the strength of evidence against a null hypothesis, with a smaller p-value indicating stronger evidence against the null hypothesis, suggesting that the observed results are unlikely due to random chance alone. Typically, we reject the null hypothesis if the p-value is below a predefined significance level (e.g., 0.05).

Second, please circle the p-value for the hypothesis test which tests the null hypothesis that $\hat{\beta}_1 = 0$.

1.49e-12 should be circled

Third, can we reject the null hypothesis that $\hat{\beta}_1 = 0$ at a significance level of 0.05?

From the model summary output, the p-value for the hypothesis test which tests the null hypothesis that $\hat{\beta}_1 = 0$ is 1.49e-12, which is much smaller than the 5% significance level. Therefore, we have strong evidence to reject the null hypothesis that $\hat{\beta}_1 = 0$.

4D (5 Points).

Suppose that a certain car's speed is 100 mph. Using the model above, how much distance would you predict that this car will need to come to a stop?

Predicted distance = -17.5791 + 3.9324*100 = 375.6609 ft

4E (5 Points).

Suppose that one car is driving 30 mph, while a second car is driving 40 mph. How much more distance will the second car need to come to a stop, compared to the first car?

Distance = -17.5791 + 3.9324*40–17.5791 + 3.9324*30 = 39.324 ft

4F (5 Points).

Please comment briefly on whether you think this is a good model, and consider particular cases (speed ranges) where we should be cautious before applying this model to make predictions.

Possible Answers:
The regression model could be a good fit for predicting the stopping distance as speed appears to be a statistically significant predictor for the stopping distance prediction. The positive relationship suggested by the model is also reasonable in the real-world context.
However, one should be careful with the eligible speed ranges when applying the model, particularly in cases involving unrealistic speed ranges (e.g., negative speeds) or speeds that significantly exceed the range of speeds observed in the data used to build the model. **Specifically, we should be cautious that the predicted stopping distance will be negative when the given speed is lower than 4.47 mph**.